

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ОРГАНА СЛУХА ДЛЯ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Ю.Н. Титов

(Тамбовский государственный университет им. Г.Р. Державина)

Научный руководитель – д.т.н., профессор А.А. Арзамасцев

(Тамбовский государственный университет им. Г.Р. Державина)

В статье дано описание модели и результаты по моделированию органа слуха при автоматическом распознавании изолированных слов русского языка. Изложен алгоритм последовательной обработки сигнала через банк фильтров с учетом психоакустической природы слуха (Mel-Scale Transform) и результаты классификации полученных векторов-признаков с помощью аппарата искусственных нейронных сетей.

Введение

В рамках задачи оптимизации интерфейса «человек–компьютер» стоит проблема обеспечения коммуникации между ЭВМ и человеком посредством голосовых команд. Интерес к распознаванию команд по голосу обусловлен тем, что существенная в процентном отношении доля прагматически важной информации в повседневном взаимодействии человека с окружающим миром, а также многие проявления его реакции на то или иное событие и явление выражается через язык посредством голоса, а восприятие акустических волн осуществляется с помощью органа слуха. Для реализации задачи автоматического распознавания речи, т.е. построения голосового интерфейса нередко используются разнообразные подходы. Проблема автоматического распознавания речи на данный момент не является до конца решенной.

Нелинейные свойства слуха

Основной проблемой при распознавании речи является выбор компактного и информативного описания речевого сигнала, при котором существенно понижалась бы размерность образа слова и при этом сохранялись основные информативные признаки, позволяющие отличить одно слово от другого.

Наиболее распространенным методом формирования цифрового представления является спектральный анализ, основанный на дискретном преобразовании Фурье. Он позволяет сгладить влияние случайной компоненты сигнала, достаточно устойчив к изменениям интенсивности (громкости) произнесения, но формируемый образ имеет большую размерность и неудобен для распознавания [7].

Наличие определенных успешных открытий в сопряженных областях, работающих над цифровой обработкой сигналов, побуждает исследователей в последнее десятилетие обращаться к подходам, которые включают анализ процессов, происходящих в биологических объектах, и дальнейший учет в разработке математических моделей, применяемых для создания тех или иных программных продуктов. Ярким примером данной тенденции служит создание стандарта компрессии аудио MPEG-2, в основе которого лежит математическая модель слухового аппарата человека с учетом физиологических особенностей восприятия акустических волн [6]. Положительные результаты в данном направлении дают основания для их применения в области цифровой обработки акустических сигналов и распознавания речи.

Наука, которая изучает восприятие акустических волн биологическими организмами, в частности, человеком, называется психоакустикой. Основные задачи психоакустики – понять, как слуховая система расшифровывает звуковой образ, установить основные соответствия между физическими стимулами и слуховыми ощущениями, и выявить, какие именно параметры звукового сигнала являются наиболее значимыми для передачи семантической (смысловой) и эстетической (эмоциональной) информации [1].

Звуковой сигнал любой природы может быть описан определенным набором физических характеристик: частота, интенсивность, длительность, временная структура, спектр и др. Им соответствуют определенные субъективные ощущения, возникающие при восприятии звуков слуховой системой: громкость, высота, тембр, биения, консонансы–диссонансы, маскировка, локализация–стереоэффект и т.п.

Слуховые ощущения связаны с физическими характеристиками неоднозначно и нелинейно, например, громкость зависит от интенсивности звука, от его частоты, от спектра и т.п. Еще в прошлом веке был установлен закон Фехнера, подтвердивший, что эта связь нелинейна: ощущения пропорциональны отношению логарифмов стимула [1]. Например, ощущения изменения громкости в первую очередь связаны с изменением логарифма интенсивности, высоты – с изменением логарифма частоты и т.д.

Улитка (cochlea) играет основную роль в слуховом восприятии. Она представляет собой трубку переменного сечения, свернутую три раза подобно хвосту змеи. В развернутом состоянии она имеет длину 3,5 см [1]. Базилярная мембрана состоит из нескольких тысяч поперечных волокон: длина 32 мм, ширина у стремечка – 0,05 мм (этот конец узкий, легкий и жесткий), у геликотремы – ширина 0,5 мм (этот конец толще и мягче). Общий механизм передачи звука упрощенно может быть представлен следующим образом: звуковые волны проходят звуковой канал и возбуждают колебания барабанной перепонки. Эти колебания через систему косточек среднего уха передаются овальному окну, которое толкает жидкость в верхнем отделе улитки (лестнице преддверия), в ней возникает импульс давления, который заставляет жидкость переливаться из верхней половины в нижнюю через барабанную лестницу и геликотрему и оказывает давление на перепонку круглого окна, вызывая при этом его смещение в сторону, противоположную движению стремечка. Движение жидкости вызывает колебания базилярной мембраны (бегущая волна).

В настоящее время на основе тщательных экспериментов, в процессе которых слушателю предъявлялись два звука разной частоты с просьбой расположить их по высоте, установлена зависимость высоты тона от частоты сигнала, показанная на рис. 1 [7].

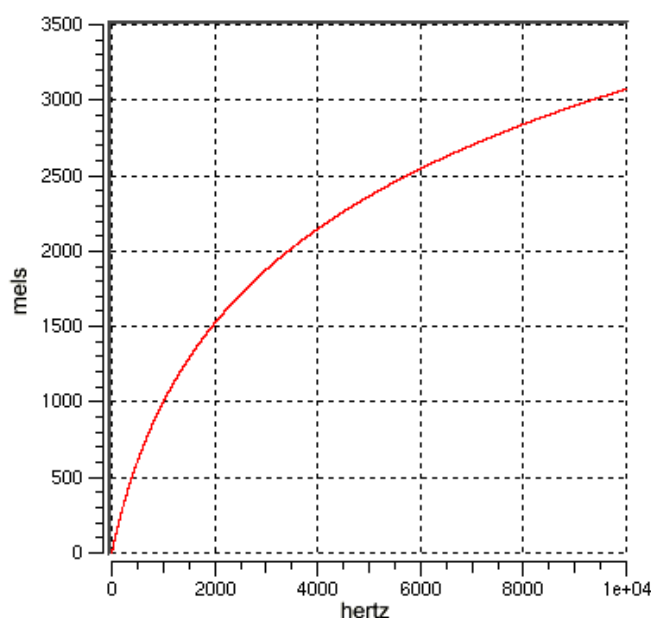


Рис. 1. Зависимость высоты тона в мелах от частоты сигнала

Значения высоты отложены в специальных единицах – мелах. Один мел равен ощущаемой высоте звука частотой 1000 Гц при уровне 40 дБ (иногда для оценки высоты тона используется другая единица, барк = 100 мел). Как видно из рис. 1, эта связь нелинейна – при увеличении частоты, например, в три раза (от 1000 до 3000 Гц), высо-

та повышается только в два раза (от 1000 до 2000 мел) [1]. Если высоту тона в мелах обозначить m , то на основе многих экспериментов зависимость тона от частоты f в герцах можно представить в виде [6]

$$m=1127,01048 \log (1+f/700). \quad (1)$$

Теория места в психоакустике при восприятии высоты основана на способности базилярной мембраны выполнять частотный анализ сложного звука, т.е. действовать как спектральный анализатор. Базилярная мембрана организована тонотопически, т.е. каждый тон имеет свою топографию размещения [1]. Как уже было указано выше, звуковой сигнал вызывает появление на мембране бегущей волны, но специфика возбуждения состоит в том, что максимум смещения этой бегущей волны располагается в разных местах базилярной мембраны – низкие частоты имеют максимум смещения вблизи вершины мембраны, высокие – вблизи овального окна. Каждая частота имеет свое место максимума возбуждения на мембране. В зависимости от спектрального состава на базилярной мембране возбуждаются различные участки. Возбуждаются волосковые клетки, находящиеся на этом месте, и их электрическая активность сообщает мозгу, какие частоты присутствуют в спектре. Таким образом, частота тона представлена в коде, основанном на том, нейроны каких участков активны, а каких – молчат.

Нейронные сети как рецепторы базилярной мембраны

Удобным инструментом для моделирования рецепторов восприятия акустических волн нейронами-рецепторами, расположенными вдоль базилярной мембраны, является аппарат искусственных нейронных сетей [4]. Развитые в последние десятилетия в связи с резким увеличением вычислительной производительности компьютерной техники, они показывают себя как весьма серьезный подход для решения задач классификации образов. Также имеет место аналогия с реальными нейронами-рецепторами – правда, с учетом некоторых ограничений, как, например, отсутствие затухания сигнала, распространяющегося в нейронной сети от нейрона к нейрону вследствие отсутствия сопротивления синапса, и т.д. [2].

При выборе числа входных нейронов и построении архитектуры нейронной сети следует учитывать некоторые фактические аспекты восприятия речи человеком. Известно, а также легко наблюдаемо из экспериментов, что диапазон используемых частот человеческой речи располагается примерно в интервале 0–4000 Гц [3]. В свою очередь, из психоакустической теории места известно, что полосам частот по пропускной способности соответствует 25 сегментов [1].

Таким образом, можно сформулировать модель слуховой системы, которая представлена из линейки 25 психоакустических (mel-scale) полосовых фильтров и нейронной сети для классификации сегментов речи человека. Для того чтобы нейронная сеть смогла произвести классификацию, необходимо сформулировать классификацию языка речи, т.е. сформировать то, что требуется классифицировать. Также необходимо создать словарь обучения, т.е. набор образцов классификации для начального обучения нейронной сети.

Эксперимент и результаты

Нейронная сеть состояла из 25 нейронов входного слоя, трех нейронов внутреннего слоя и одного выходного нейрона. Использовалась сигмоидальная функция активации. В эксперименте использовался речевой корпус из десяти числительных, а также набор сэмплов, которые представляли собой названия 40 городов РФ. Численность дикторов составляла 17 женщин и 15 мужчин. Для разделения фонем и выделения участков одной фонемы с определением границ для дальнейшей обработки использовался

алгоритм кэпстра мел-скейл преобразования, описанный в [8]. Результаты представлены в табл. 1.

Степень распознавания, %	Дикторы	
	Мужчины	Женщины
Числительные	85	87
Названия городов	71	73

Таблица 1. Результаты распознавания изолированных слов из речевого корпуса

Заключение

В статье дано описание математической модели среднего уха человека с помощью психоакустического подхода восприятия высоты и полученной с помощью него классификации образов с использованием аппарата искусственных нейронных сетей. Приведены результаты экспериментов распознавания изолированных слов. К достоинствам данного метода можно отнести его достаточную простоту реализации, а также весьма очевидную аналогию с процессами, происходящими в реальном органе слуха человека. Недостатком является уровень ошибки при распознавании (13–23 %), который предлагается снизить использованием блоков контекстного распознавания, таких, как например, описанных в [9]. Также предлагается использовать увеличение речевого корпуса для обучения, так как большие речевые корпуса при обучении нейронных сетей играют важную роль для правильной классификации, представляя статистический материал для обучения и последующего распознавания [5].

Литература

1. Алдошина И.А. Основы психоакустики. //Звукорежиссер, 2002. №3. С. 86–92.
2. Барский А.Б. Нейронные сети: Распознавание, управление, принятие решений. М.: Финансы и статистика, 2004.
3. Кривнова О.Ф., Захаров Л.М., Строкин Г.С. Речевые корпуса (опыт разработки и использование) / Диалог'2000. Прикладные проблемы. М., МГУ им. М.В. Ломоносова, 2000.
4. Леонович А.А., Медведев М.С. Распознавание фонем: функциональный и нейросетевой подходы. / Материалы XXIII международной конференции «Информационные технологии в науке, образовании, телекоммуникации и бизнесе», Красноярск, КГТУ, майская сессия, 2003.
5. Лобанов Б.М., Цирульник Л.И. Фонетико-акустическая база данных для многоязычного синтеза речи по тексту на славянских языках.
6. Салюшин С.А. Методика экспериментального определения структурных параметров нейросети для распознавания речи. / Сборник научных трудов МИФИ, 2004.
7. Федяев О.И., Гладунов С.А. Фонетический анализ речи на основе нейросетевой аппроксимации сигнала. / Научные труды Донецкого гос. технического университета. Серия: Проблемы моделирования и автоматизации проектирования динамических систем. 1999. С. 38–43.
8. Dusan S., Rabiner L. On the Relation between Maximum Spectral Transition Positions and Phone Boundaries. // IEEE proceedings, Rutgers University, Piscataway, New Jersey, USA, 2001.
9. Cummins F., Grimaldi F., Leonard T. The Chains corpus: characterizing individual speakers. / Proceedings of the 11th International Conference «Speech and Computer SPECOM'2006». St. Peterburg, Anatolya Publishers, 2006.